The Citi Bike data

New York City launched its bike-sharing program, Citi Bike, in 2013. To date the program has seen over 20 million rides from individuals of all ages, in all weather conditions, at all points of the year.

New York City hosts records of all the Citi Bike rides in NYC on its Open Data platform. The records include the day and time of the ride, the age and gender of the rider, the start and stop stations of the ride, the unique bike ID, and the trip duration.

Using this data, we examined whether it is possible calculate the rate at which riders slow down as they age. Given that we can identify bike trips that follow the same path (from station A to station B) and leave at nearly the same time (within an hour) by age, there is the potential to build out a curve of "slowdown by age".

Ultimately we narrowed our data down to 1.3 million rides for which there are multiple people of different ages who followed the same path within an hour of each other. However, as will be discussed below, our results are by and large inconclusive.

One potential explanation is that age is not the determining factor of how fast a Citi Bike rider goes. While our initial exploration and modeling suggests that age almost certainly has some effect, the more significant determinant is likely traffic, both pedestrian and vehicular. A biker can only travel so fast in Times Square during rush hour, no matter the age. This effect is compounded since Citi Bike traffic is most common during standard commuting hours.

Due to the lackluster results from our initial analysis, we also compared the Citi Bike biking times to the expected biking time as predicted by Google Maps, by age. In this case, the initial results appear promising.

ETL & Parallelization

The data were stored in a Postgres Database via Amazon Web Services. Due to the cleanliness of the data provided in NYC Open Data, this solution provided rapid access and manipulation after some limited initial processing and calculations. Earlier attempts to process the data in Hive and SparkSQL failed due to the slowness of manipulating the data.

Data manipulation and graphing were conducted in RStudio Server, using the RPostgresSQL, dplyr, ggplot2, and multiplyr packages. Multiplyr allowed for parallelization of the computations to 7 cores, drastically speeding up the process.

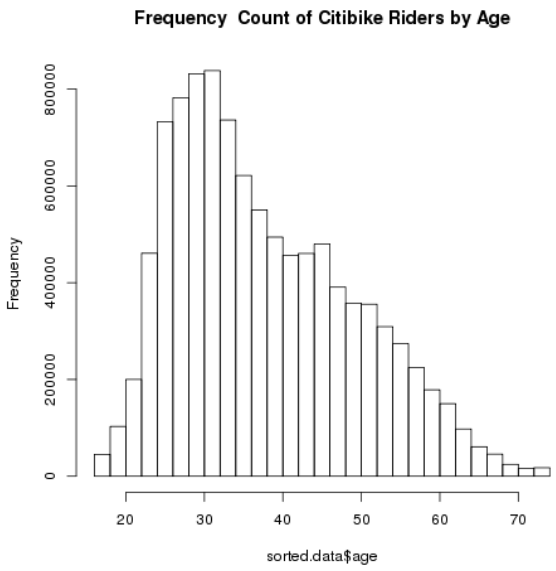
Google Maps data were downloaded using the ggmaps package for R, and the results were stored in a CSV file on Amazon Web Services and then subsequently re-loaded into RStudio for processing.

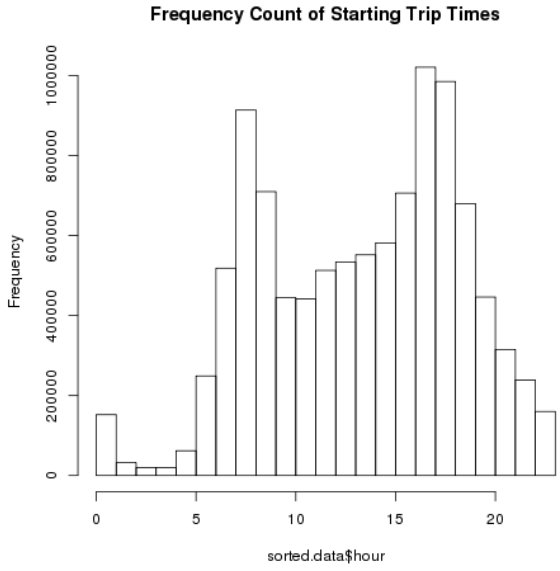
Data Description

The total dataset contained 24 million rides between May 2013 and February 2016. From this data, we removed all rides for which the stop station and the end station were the same (indicating a round trip), and analyzed the results.

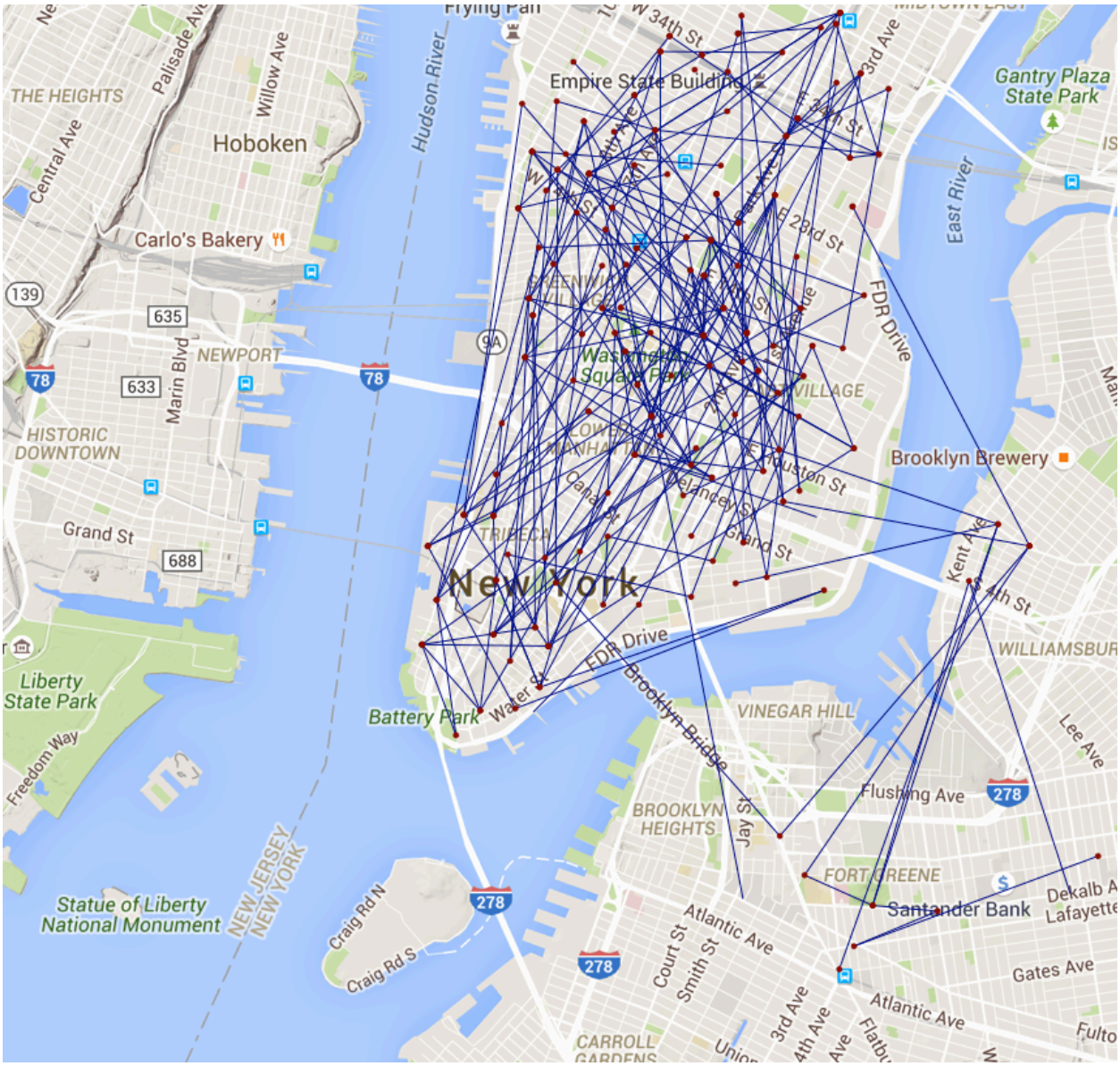
The median age of a Citi Bike rider is 36 years old, with the peak of the distribution between 28-32 years old. The bulk of the rides occur at rush hour, suggesting that many riders use Citi Bike as a commuter

device. The median time is only 10.3 minutes, suggesting that most Citi Bike riders tend to travel to nearby destinations.





We also explored the data by looking at the paths of a single bike during the course of a month, specifically May 2014. This data is mapped out below. The red dots indicate stations where at least one trip commenced.



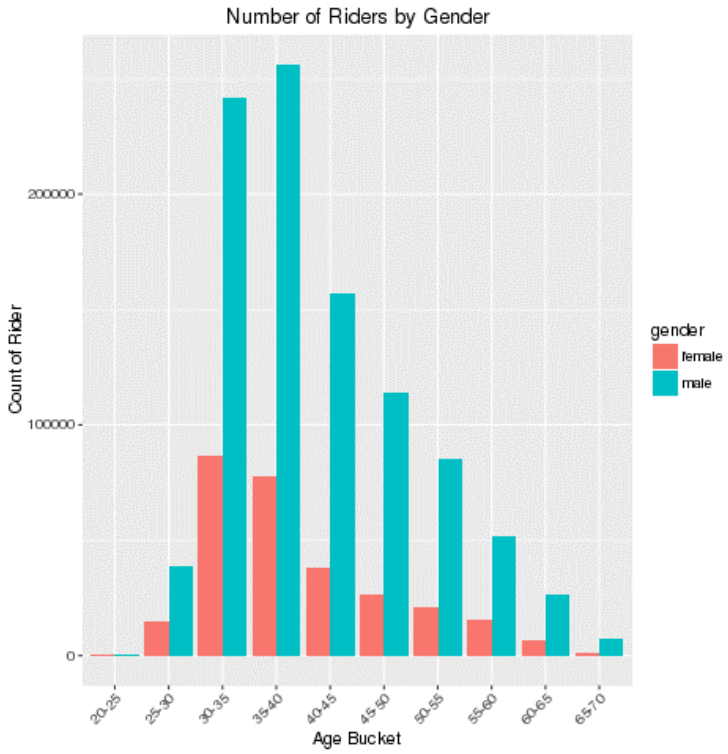
This map shows that the majority of Citi Bike trips occur in lower Manhattan, though there is some activity into and around Brooklyn. This makes sense based on our research of the Citi Bike expansion. Citi Bike grew significantly into Queens and upper Manhattan later on. During 2014 when this data collected, the program it was limited to the area covered on this map. We created the map using the maps package in R.

Data Processing

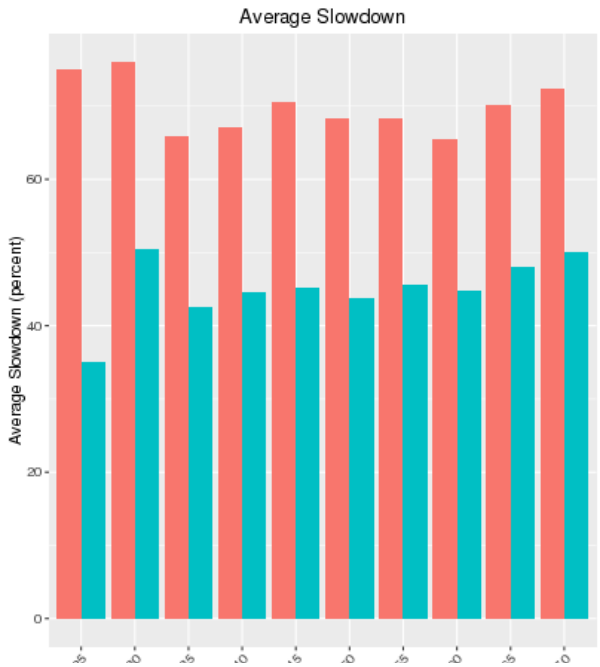
To ensure that weather, traffic and other confounding variables did not bias our results, we narrowed the data set to rides that occurred along the same path within the same hour (defined as hour on a clock, i.e. between 8 am and 9 am, not on a rolling 60 minute basis).

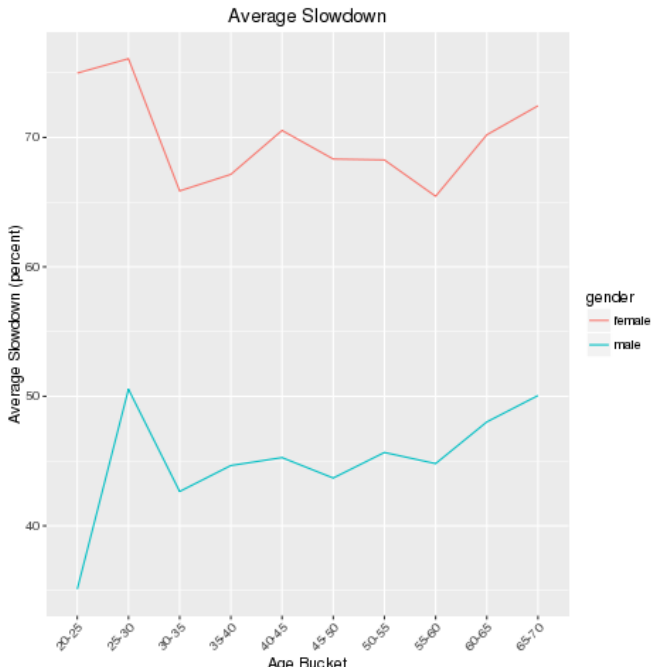
We then divided our ridership into 5-year age buckets, beginning with ages 15-20 and ending with ages 70-74. After filtering the data in this way, we retained approximately 900,000 rides concentrated among

men ages 30-35 and 35-39. This age and gender split is generally representative of the data as a whole, with the exception that there are disproportionately fewer older riders (older than 60-65) than in the full data.



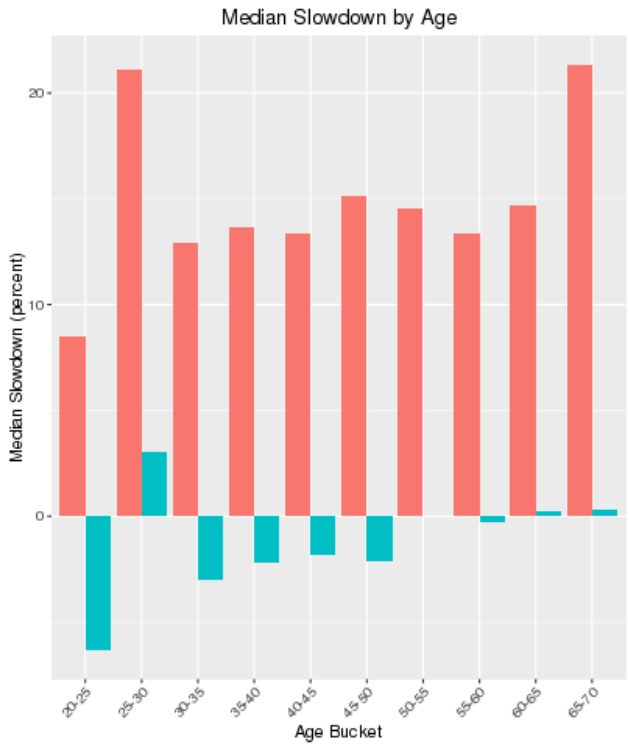
Once the processing was completed, our initial results showed that, on average, men slowed down by roughly 40-50% for every 5 years of aging, while women slowed by 60-70%. Unfortunately, this result does not make sense.

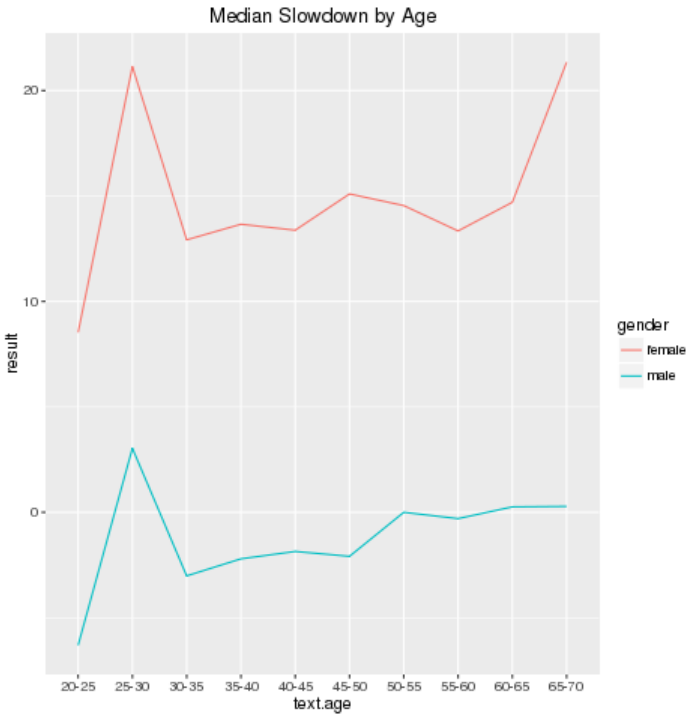




The data show that as people age, we tend to slow down by roughly the same percentage every 5 years. This conclusion is generally unsatisfactory. It largely defies common sense to believe that a nearly identical level of reduction in physical capacity accompanies aging from 25 to 30 and from 40 to 45.

A look at the median slowdowns similarly confirms that our data do not conform to simple common sense. The medians show that the median man actually modestly speeds up as he gets older. Perhaps the large acceleration (negative slowdown) at ages 30-35 is caused by selection bias: only athletically active men continue to use Citi Bike after the age of 30. However that explanation is taken to its extreme by the time riders are between 45 and 50. We would have to believe that men in that range are traveling faster than your average, already athletic 30-year-old.

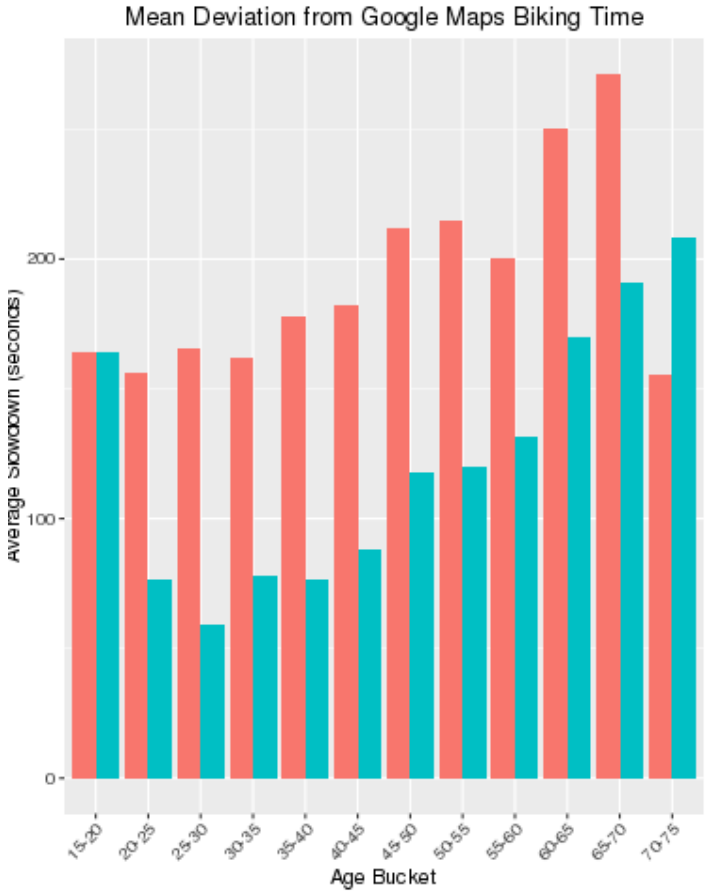


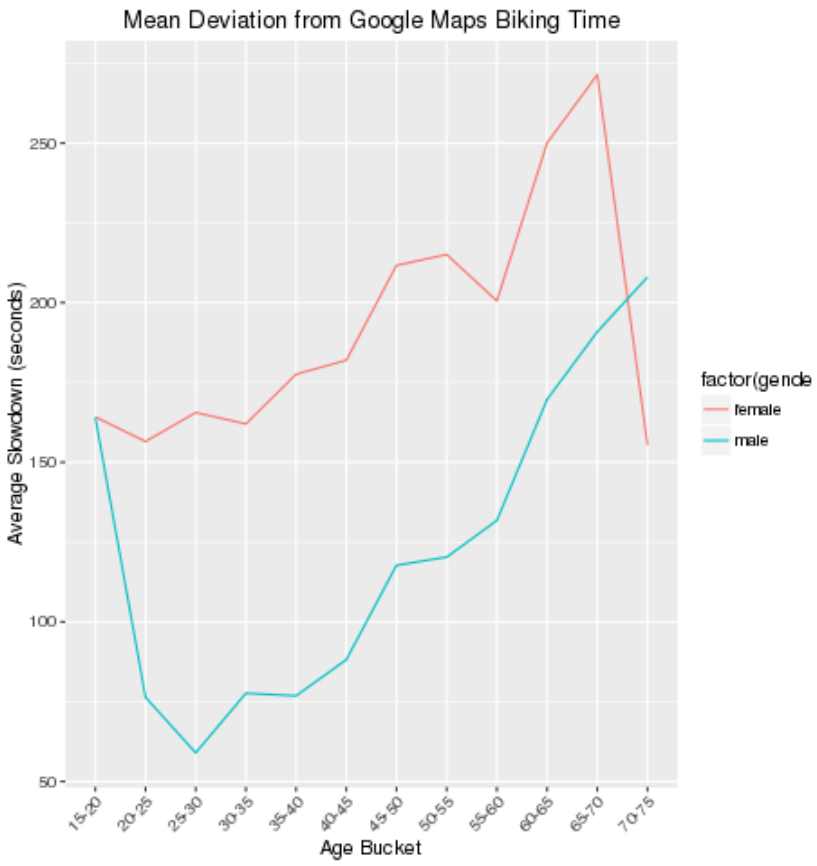


Google Maps Strategy

Due to the disappointing results of our first analysis attempt, we attempted another strategy, supplementing with data from Google Maps. Google Maps allows users to download the forecasted biking time between two locations in its database. Users are allowed to download 2,500 free data points, though only half of those if geocodes are used, per day.

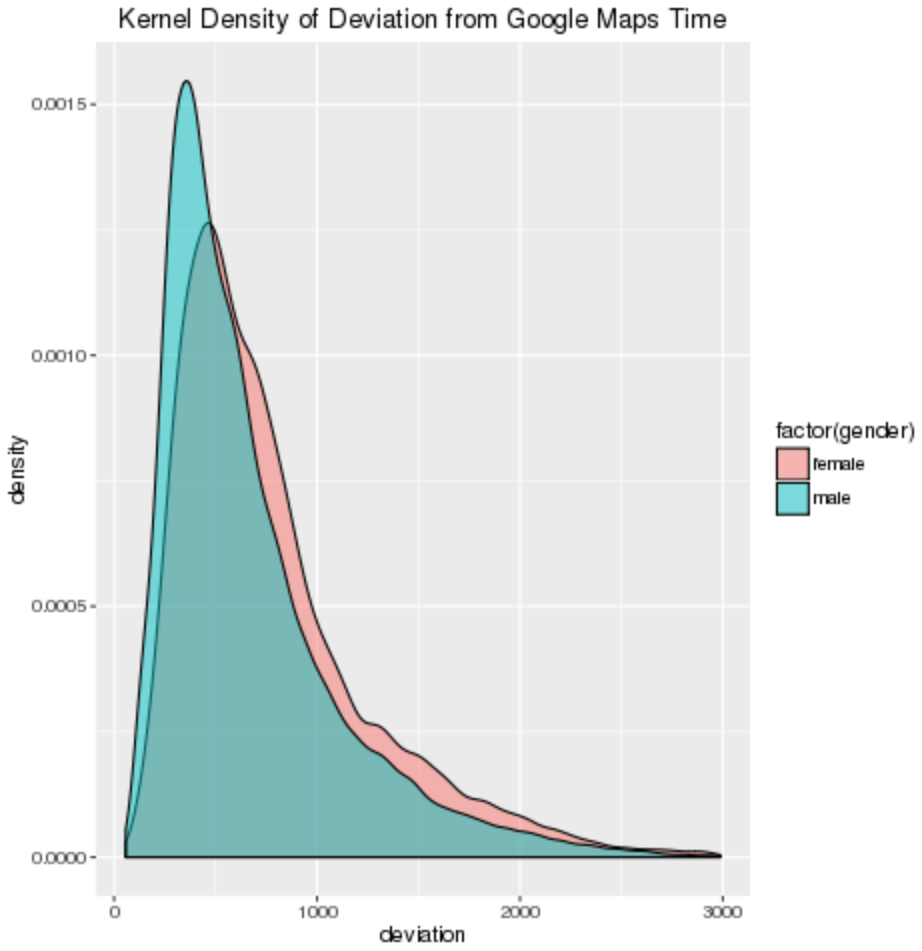
For our analysis, we generated 124,000 unique pairs of stations and randomly sampled from those the key pairs we would analyze. We calculated the average *deviation* of Citi Bike's trip duration from Google's expected time.





These results are far more in line with intuition. These results show a continual slowdown from ages 25-30 all the way until 65-70. Quirks in the data, including the dramatic increase in speed deviation in women in the 70-75 bucket, are likely due to pure randomness and should likely be discarded. Similarly, men appear to speed up between the 15-20 year-old bucket and the 20-25 year-old bucket. Several explanations are possible: perhaps the younger cohort represents teenagers who are riding with their parents, and as a result are traveling slower. Another possibility is selection bias: among riders age 20-25 we are seeing athletic workers choosing to bike on their commute, while the younger set is more representative of the entire population.

This analysis suggests two other results. First, Google Maps biking time appears to underestimate the time it takes to bike between two locations on Citi Bike – either due to incorrect assumptions by Google, or, much more likely, the fact that Citi Bike bikes are particularly heavy bicycles. Second, Google Maps data appear to be much closer to the expected biking time for men than for women.



Scalability

The scalability obstacles to expanding this project are low. The full database increases at a rate of about 600K-700K records per month, which is large but certainly not prohibitive. The current set-up in Postgres could easily support that volume, provided that the AWS storage is sufficiently large.

The biggest change in terms of scaling our analysis would probably be a change in our analysis software. R is great for exploration, which was our first priority in this project. By adding parallelization, we were able to dramatically speed up the exploration process. However, our scripts are not optimized for quick processing – using data.table rather than dplyr would further improve speed, for instance. If the system were to be productionalized into a tool or real-time app, R wouldn't be the right tool.

Future Research Directions

Though the initial results from the Google Maps analysis are promising, there are many additional ways to explore this dataset further. We were unable to incorporate weather or traffic into our final analysis, so a follow-up project could return to the initial analysis plan and calculate differentials. Part of the reason we abandoned that approach was because it was too difficult to get effective benchmarks for

path timing at the hourly level – there were over 150,000 potential paths, and even with 24M records, there was less path overlap between rides than we were expecting. By adding in weather data from NOAA and traffic data from the NYC Department of Transportation (perhaps using vehicle crashes or traffic advisories as a proxy), we could more effectively distinguish the impact that those factors had on biking speed from the effect of age.

We are interested in other potential uses of this dataset as well. A 2015 study from Resources for the Future suggested that the Capital Bikeshare program in DC reduced congestion in surrounding areas (defined by census blocks) by 2-3%. By combining the data with the aforementioned DOT data, it would be interesting to see if that result could be substantiated.